Deep learning for prediction of future endoscopic disease activity in Ulcerative Colitis

R. Oniga, M. Blaschko, T. Eelbode, F. Maes[a], Raf Bisschops[b], Peter Bossuyt[c]

[a]*Department of Electrical Engineering, ESAT/PSI, KU Leuven, Leuven, Belgium,*
Medical Imaging Research Center, UZ Leuven, Leuven, Belgium
[b]*) Department of Gastroenterology and Hepatology, UZ Leuven, Leuven, Belgium,*
Department of Translational Research in Gastrointestinal Diseases (TARGID), KU Leuven, Leuven, Belgium
[c]*Imelda Ziekenhuis, Bonheiden*

Email addresses:
oniga.robi@gmail.com, matthew.blaschko@kuleuven.be, tom.eelbode@kuleuven.be, frederik.maes@kuleuven.be (R. Oniga, M. Blaschko, T. Eelbode, F. Maes), raf.bisschops@uzleuven.be (Raf Bisschops), peter.bossuyt@imelda.be (Peter Bossuyt)

Abstract

Ulcerative Colitis is an inflammatory bowel disease that affects the lower gastrointestinal tract which is composed of the colon and the rectum. The disease exhibits itself with alternating periods of acute phases and remission during which the patient can suffer various clinical manifestations. The best way at this time to asses the disease is through colonoscopies where the clinician looks for clinical symptoms such as redness, ulcerations, bleeding, stool frequency all of these being part of a scoring system called the MAYO scoring system. The main limitation of this scoring system is the high subjectivity of the clinician that takes part in the assessment of the disease. This calls for an automated method of both diagnosing and scoring the disease using machine learning algorithms that are capable of detecting even the slightest differences in the evolution of the disease such that the treatment of said patient can be adjusted accordingly while predicting the clinical outcome: remission or non-remission.

Keywords: *Ulcerative Colitis, deep learning, segmentation, scoring, clinical outcome, CNN, MAYO score.*

1. Introduction

At a global level, many persons are affected by Inflammatory Bowel Diseases (IBD), however, it was concluded that in the Westernized nations, the prevalence of this types of diseases is the highest, with the incidence ranging from 0.6 to 24.3 per 100,000 in Europe, 0 to 19.2 per 100,000 in North America, and 0.1 to 6.3 per 100,000 in the Middle East and Asia [1]. Inflammatory bowel diseases which comprise of Crohn's disease (CD) and Ulcerative Colitis (UC) are chronic immunologically mediated diseases that appear due to dysregulations in the immune response to a normal or altered gut microbiome in subjects that present some genetical susceptibility. IBD has a high impact on the quality of life due to the early onset characterized by alternating periods of remissions and relapses, and they contribute to a significant morbidity at a global level [2].

Ulcerative Colitis affects the mucosa of the large intestine, most usually at the level of rectum and the sigmoid colon. At a histological level, between the villi in the lining of the intestinal epithelium there are glands also called crypt of Lieberkuhn which become inflamed causing release of cytokines activating macrophages. These contribute to the onset of the acute phase of the disease damaging the epithelial mucosal barrier which causes leak of fluids into the gut. These are the microscopical manifestations of the disease. Macroscopically speaking, depending on the degree of inflammation, ulcerative colitis has

ISSN

Characteristic	Presentation	Score
Stool frequency	Normal	0
	1-2 stools/day more than normal	1
	3-4 stools/day more than normal	2
	>4 stools/day more than normal	3
Rectal bleeding	None	0
	Visible blood with stool less than half the time	1
	Visible blood with stool half the time	2
	Passing blood alone	3
Mucosal appearance at endoscopy	Normal or inactive disease	0
	Mild disease (decreased vascular pattern)	1
	Moderate disease (erosions, absent vascular pattern)	2
	Severe disease (spontaneous bleeding, ulceration)	3
Physician rating of disease activity	Normal	0
	Mild	1
	Moderate	2
	Severe	3

Table 1: MAYO scoring system.

consequences on the bowel that present themselves as dark red or velvety patches on the inner layer of the intestine or in worse cases, hemorrhagic patches that turn into ulcers.

The most important factor of the diagnosis procedure is the imaging technique which consists of endoscopic investigation that allows the clinician to visually analyze the entire colon. Also, this process allows the clinician to determine the severity of the inflammation along with the extent of the disease. The imaging procedure is called a lower GI endoscopy or colonoscopy which allows the medical practitioner to examine the entire colon and rectum. During the procedure, a long, flexible tube that is fitted with a camera is inserted into the rectum and travels through the entire colon so that it can be thoroughly examined. The device blows air into the intestine so that it expands, giving the doctor a clearer view of the area. In some cases, if abnormalities are present such as growths, or polyps, small amounts of them can be taken out with the help of a biopsy clamp that pinches part of said abnormality in order to be further examined outside of the body.

In patients suffering from Ulcerative Colitis, during the colonoscopy, the physician grades the severity of the disease using one of the many existing scoring systems. The most popular scoring system is the MAYO scoring system also called the Disease Activity Index. This system evaluates four aspects of the disease: stool frequency, rectal bleeding, mucosal appearance at the endoscopy, and the rating given by the physician based on his observations. Each of these characteristics have a way of presenting themselves which was summarised in Figure 1. The total score ranges from 0 to 12. A total score of 3 to 5 points indicates mild disease activity, 6 to 10 moderate disease activity, and 11 to 12 severe disease activity.

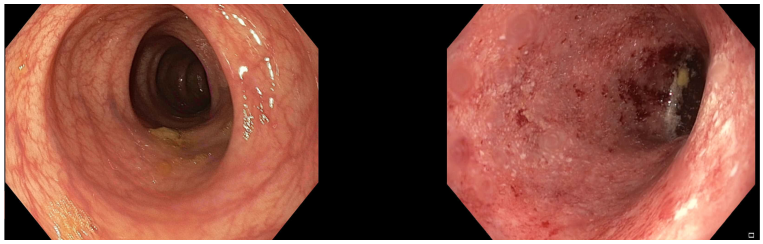

Figure 1: Example of healthy vs inflamed colon.

An overview of recent research that has been done in the field of IBD is presented by John Gubatan et.al. [4] which summarizes the efforts made in diagnosing and assessing inflammatory bowel diseases

(both ulcerative colitis and Crohn's disease) by combining deep learning and artificial intelligence models. Most attention was given towards diagnosis and risk prediction, then the evaluation of the disease activity, and the least amount of interest was directed towards delivering a clinical outcome. When it comes to the AI methods used in order to overcome the problems that each group of researchers has tackled, the most preferred ones were neural networks (both deep and convolutional), RF (random forest), and SVM (support vector machines). Since the exact onset factor of the disease is unknown, multiple data modalities have been chosen in order to build the AI model that would yield the desired results. Many studies (16/22 that focused on diagnosis and risk prediction) used genetic/genomic datasets, only 4/58 used imaging or endoscopic datasets, and the least interest was given to expression/proteomics datasets with only 2 studies. The assessment of disease activity and grading of the severity of Ulcerative Colitis was done using validated clinical scores (Mayo scoring system) which was subjected to recall bias, heterogeneity, intraobserver and interobserver variability. Here both endoscopic (11 out of 13 studies) and histologic (2 out of 13) datasets have been used in order to monitor inflammation. Finally, the studies oriented towards delivering a clinical outcome used molecular datasets (4 out of 16 studies), electronic health records (11 out of 16 studies), and histologic datasets (1 out of 16 studies). Out of all the methods that were employed, the ones that yielded the best accuracy, sensitivity, specificity, for any of the three applications (diagnosis, disease activity grading, clinical outcome) were SVM, however CNN had the most prominent growth potential along the years fact that was observed in this article [4]. *The focus of this work is mainly towards studies that use endoscopic images obtained through colonoscopies in order to detect inflammations, measure severity of the disease, and use both to predict the evolution of Ulcerative Colitis.*

2. Methods

In order to successfully complete the goal of this project, the work carried out during the research project was split into seven parts: obtaining the data (get the recordings taken during the colonoscopy procedures), data structuring (create a hierarchical structure and format all data to be identical from an encoding point of view), frame selection (automate the process through which medically relevant frames are taken from the recordings in order to use them for training a network to recognise inflammation), annotation (create masks that show the exact location of the inflammation in each selected image), delineation (create an algorithm that is able to distinguish between healthy and inflamed tissue), scoring (come up with a system that is able to measure the degree of inflammation), and finally prediction of clinical outcome (based on the predictions made by the network, determine how the disease will develop in time). This structure is summarized in Figure 2. The data was obtained from a clinical trial (LOVE_UC) that took place between 2015 and 2020 in Belgium, Netherlands, and Hungary where patient with Ulcerative Colitis had a series of colonoscopies: at the time of screening (week 0)„ week 26, and week 52 where the progress of the disease was measured using the MAYO scoring system while the patient was on treatment using Vedolizumab. The data consisted of videos of colonoscopies and the data was different depending on the center at which the images were acquired due to the use of different endoscopic equipment. The data was organized hierarchically with a naming system so that patients could be easily identified along with the period of time from which the recording came from (screening, week 26, week 52). Afterwards, in order to build a strong model, frames were automatically selected making sure that the whole spectrum of the disease is present in the final data set. Even though the frame rate was 25 FPS, it was clearly seen by visual analysis of the recordings, that one frame per second is more than enough to be taken into consideration since the speed at which the endoscope moves through the colon is not high enough so that the field of view changes that fast. Based on this observation, one frame per second was taken from a video and by using a Sobel operator, the edges were detected in each image. The higher the number of edges, the more sharp the image, so more medically relevant information could have been taken from that image. The best 10% frames were selected in terms of sharpness for each video and this ended up creating the dataset which had a total of 1000

frames after discarding some additional ones containing biopsy needles or stool. These were then sent to a clinician whom had to use the annotation tool in order to mark the inflamed regions from each selected frame. Using the information provided through the annotation tool by the expert endoscopist, masks were generated for all 1000 frames which were then used in the training of a deep neural network which is intended to detect frames with inflammation and also segment the inflamed regions: U-NET Xception style model. Out of the 1000 images, 200 were used as independent test set and the other 800 were used for trainingusing 5-fold cross-validation (640 images in the training set, 160 in the validation set). In order to make sure that the network is performing accordingly, an ImageDataGenerator is used to feed batches of 8 images to the network, making sure that the images are augmented along with their corresponding masks. The simulation runs for 500 epochs which was empirically discovered to be optimal since the loss function did not further decrease and at the same time it did not overfit. After the network was successfully trained, a custom scoring method was developed in order to be able to predict the clinical outcome of each patient. The scoring represents the ratio of inflamed pixels at a frame level, and an average of the ratio at a video level. Finally, using a classifier, the threshold that yielded the best performance in terms of specificity and sensitivity was determined in order to differentiate between patients that were or not in remission. When we talk about sensitivity and specificity we refer to the true positive rate or false positive rate of correctly identifying the pixels from an image (i.e. whether or not a pixel is correctly identified as inflamed or not).

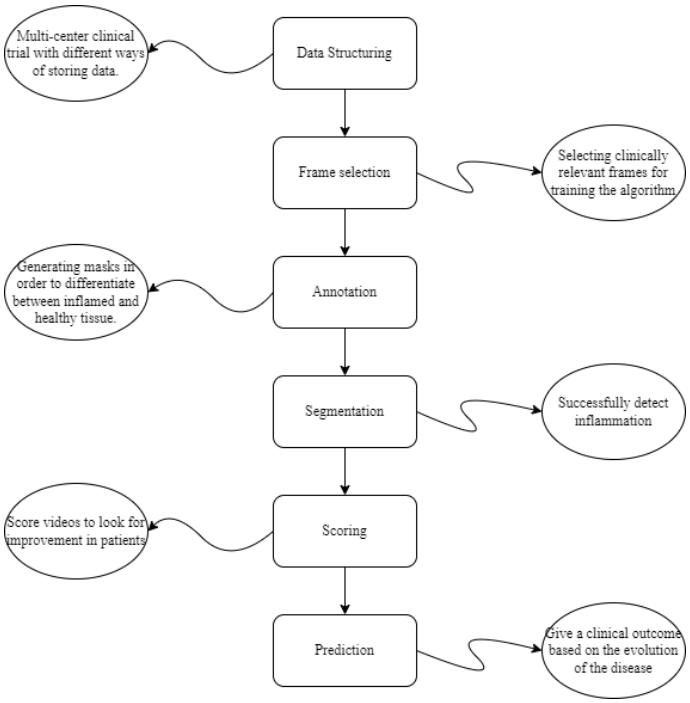

Figure 2: Flowchart of the methodology.

3. Results

The performance of the model can be seen in Table 2. The network performs slightly better on the validation set then on the training one since the validation set was not augmented in any way, making it easier for the network to predict the frames. The performance on the test set set was lower than on the training and validation set as can be seen in Table 2.

The next step was to score each of these predicted frames and compare the scores with the assigned MAYO scores in order to see how the network is able to perform. It was clear that differentiating between inflamed frames and healthy ones was not a challenge for the network, but the differentiation between MAYO 2 and MAYO 3 was nearly impossible. This was clear since the difference between the two classes

Metric	Av training set	Av validation set	Av test set
Accuracy	91%	94%	78%
Sensitivity	89%	94%	76%
Specificity	94%	95%	81%
Dice score	93%	94%	80%

Table 2: Performance metrics for the training and validation set.

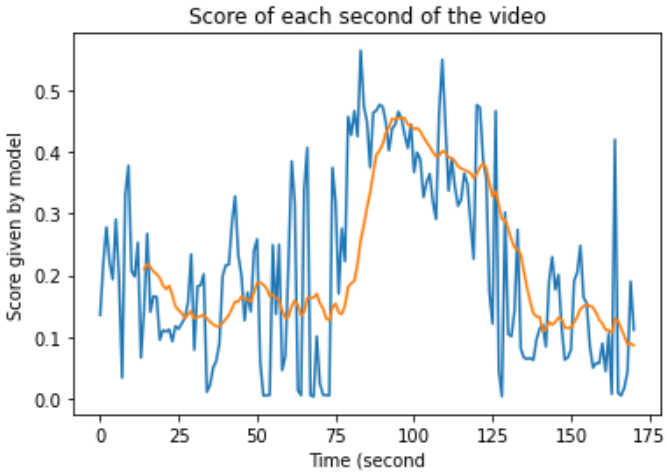

Figure 3: Scores given by the system for each second of a video.

is purely subjective, due to the score assigned by the grading physician based on his observations. When it comes to video level scoring, thing are getting more complicated, and the scores do not correlate with the MAYO scores and the differentiation between remission and non remission patient cannot be made. The scores assigned by the network at a video level over the length of a recording can be seen in Figure 3. Unfortunately, the high score region of the recording does not completely correlate with the presence of inflammation. The system assigned a high score in this particular case where only the biopsy needle was present and almost no inflammation.

At a frame level the difference between remission and non remission is clear and it can be consulted in Figure 4, and Figure 5, respectively.

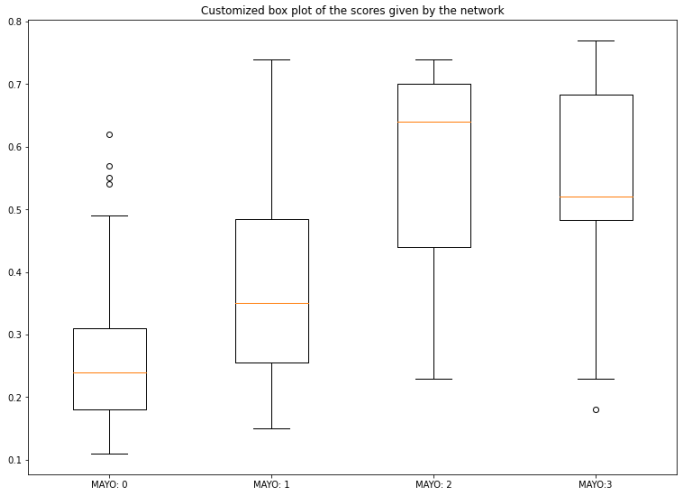

Figure 4: Scores given by the system correlated with the MAYO scores.

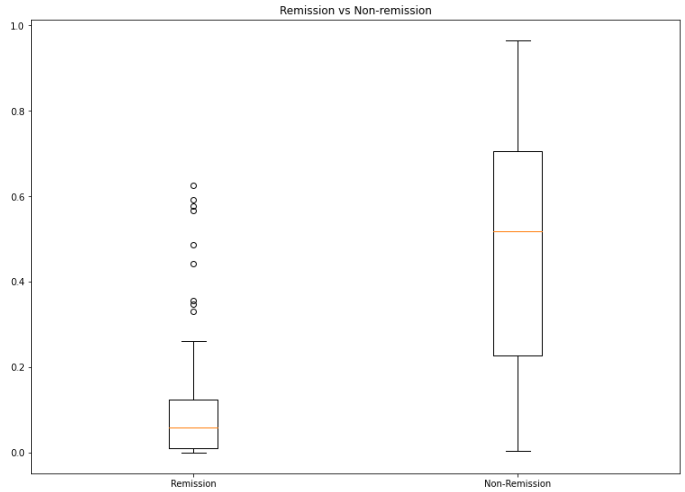

Figure 5: Remission versus Non-Remission frames scored by the network.

4. Conclusion

The main condition that had to be attained in order to successfully fulfill the goal of this project was to detect inflammation with a high specificity and sensitivity. This task was completed using 1000 frames that were manually annotated by a clinician and the results exceeded the expectation since the dataset was highly unbalanced. The majority of the frames used in the training process were completely healthy, a more significant fraction of the remaining frames were fully inflamed and lastly, only a few partially inflamed. It was anticipated that the imbalance will cause difficulties for the network to learn those characteristics that best describe inflammation, but in the end it turned out that regardless of how the data was distributed, the network built during this thesis work was good enough to detect with a *specificity* of *0.95* those frames that were used in the validation set and with a *specificity* of *0.81* the frames that were kept separately in the test set. While the results represent a positive outcome of the research, the highly unbalanced dataset represents a limitation of this work. In future work, it should be assured that each type of frames are present in the training set so that the network is capable of learning all possible instances of the disease, of how it may present itself.

After the successful completion of the segmentation task, a custom scoring method was created with the aim of measuring the differences between patients that could be considered to be in remission and those that could not. The goal of the thesis was to predict the clinical aspect of the disease, in other words whether the patient can or cannot be considered to be in remission. The scoring at a frame level represented no challenge as it was clear whether there was or not any inflammation present in them. By constantly keeping track of performance metrics such as sensitivity, specificity and area under the curve, the threshold that differentiates between patients in remission and patients that were still in acute phase of Ulcerative Colitis was found. By using this threshold, it was easy to predict the clinical outcome with high certainty, having a *specificity* of *0.85*. Even though the custom scoring system was not able to show differences between frames classified MAYO 2 or MAYO 3, it was still considered a good system since it was capable of differentiating between MAYO 0 and the rest of the classes which in clinical terms means remission and non-remission. The limitation of this scoring system is that it works best just in those instances where the interest is at a frame level only. It was clear that when working with lengthy videos, the scoring system fails to quantify the severity of Ulcerative Colitis due to the large number of healthy frames. Clinically speaking, Ulcerative Colitis will always present itself continuously, and most likely only in one segment of the colon. The videos used in this thesis were not broken up in colonic segments which clearly lowers the final score which is attributed to the video, thus "fooling" the algorithm to think that the patient is in remission even though it is not the case. In the future, a more precise algorithm has to be developed, one that works only on one segment of the colon at a time, and which is capable of only taking into consideration those frames that have inflammation

present, otherwise the video itself has to be considered part of the remission class.

Bibliography

[1] Dennis L. Kasper, Harrison's Gastroenterology and Hepatology, McGraw Hill Education, 2017.

[2] Russell D. Cohen, Inflammatory Bowel Disease Diagnosis and Therapeutics, Humana Press, 2016.

[3] Kathryn L. McCance and Sue E. Huether, Pathophysiology The Biologic Basis for Disease in Adults and Children, Mosby Elsevier, 2019.

[4] John Gubatan, Steven Levitte, Akshar Patel, Tatiana Balabanis, Mike T Wei, Sidhartha R Sinha, Artificial intelligence applications in inflammatory bowel disease: Emerging technologies and future directions, World journal of gastroenterology 27(17):1920, 2021.